Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

# 叙词表多表联合标注系统设计与实现*

李鹏 朱礼军 刘亚洁 / 中国科学技术信息研究所 北京 100038
辛之海 / 开源旗帜软件（北京）有限公司 北京 100125

摘要：通用叙词表提供了普遍意义的概念，具有普适性、协调性与兼容性的特点，而不同的专业叙词表提供了领域内关注的不同侧重点。叙词表多表联合标注能够从多视角下揭示文档的语义。文章提出了叙词表多表联合标注系统的设计方案，介绍了叙词表文本标注等功能模块以及设计中应该注意的问题，并以皮肤病领域下文档标注为例进行了研讨，总结了多表联合标注可能的应用场景。多表联合标注系统为挖掘不同视角下文档的意义提供了参考，并为文档的语义检索奠定了基础。

关键词：叙词表，标注，语义
DOI：10.3772/j.issn.1673—2286.2013.11.005

## 1 引言

标引是对文献的内容及其他有检索意义的特征进行分析、描述并用检索标识记录下来，作为存取依据的文献处理过程，它又分为分类标引和主题标引。标引不仅是信息过滤的必要组成部分，也是对原信息的精炼与提升，可以使检索更有效率，更为精准。曾经一度，因为全文索引逐渐被人采用，对于标引的需求下降，导致研究减少。但随着研究的深入，尤其是面对海量信息的检索与挖掘，如自动摘要、文本分析、主题检索等的需要，标引显得愈发重要[1]。

通用叙词表提供了普遍意义的概念，具有普适性、协调性与兼容性的特点，而不同的专业叙词表提供了领域内关注的不同侧重点。单表标引相对作用有限，但是叙词表多表联合标注能够从多视角下揭示文档的语义。例如，将一个表示事物的叙词和另一个表示该事物某个属性或某个方面的叙词所进行的联合标引，其结果可以形成一个专指概念。例如："信号模拟器稳定性"可用"信号模拟器"与"稳定性"组配，即用事物及其性质来表达专指概念。

在自动标引方面，李素建等[2]利用最大熵模型进行自动标引的研究，通过建立最大熵模型的特征集合，提出分类试验、正例试验、打分试验三种试验，总结了最大熵模型的优点在于可以灵活地选择各种特征，结合大量的特征到模型中去。章成志[3]为了有效利用标引对象的特征，并考虑到抽词标引可以转换为序列标注问题，提出基于条件随机场的自动抽词标引模型，认为是到目前为止解决序列标注问题的最好方法。程传鹏[4]针对微博文本的特点，根据微博文本中的名词或动词之间语义相似度构造图的邻接矩阵，再利用Pagerank算法思想来计算词语的重要度，作为标引词选择依据。利用叙词表进行机辅标注方面，杨贺[5]等基于海量文献人工标引，运用计量分析法对多年来积累的人工标引词从词频、词长、词类型、词共现等多方面进行分析，运用字面相似度计算词间关系来建立适用于机标和后控词表的自然语言词表的过程。朱嘉贤等[6]为支持Web资源内部信息的检索，研究多粒度语义标注，即按树根结点、分支结点、叶子结点及资源信息元为粒度单位对Web资源进行组织管理，并在此基础上探讨基于本体的搜索技术。

从以上关注标注方面的研究和设计可知，标注是一个相对主观和灵活的行为。本文主要着重相关实现，包括自动标注与手工标注，提出了叙词表多表联合标注系统的设计方案。

## 2 多表联合标注系统整体设计

本系统采用的自动标注是词表

* 本文系国家"十二五"科技支撑计划项目"科技知识组织体系的协同工作系统和辅助工具开发"（编号：2011BAH10B02）和"面向外文科技知识组织体系的大规模语义计算关键技术研究"（编号：2011BAH10B04）的研究成果之一。

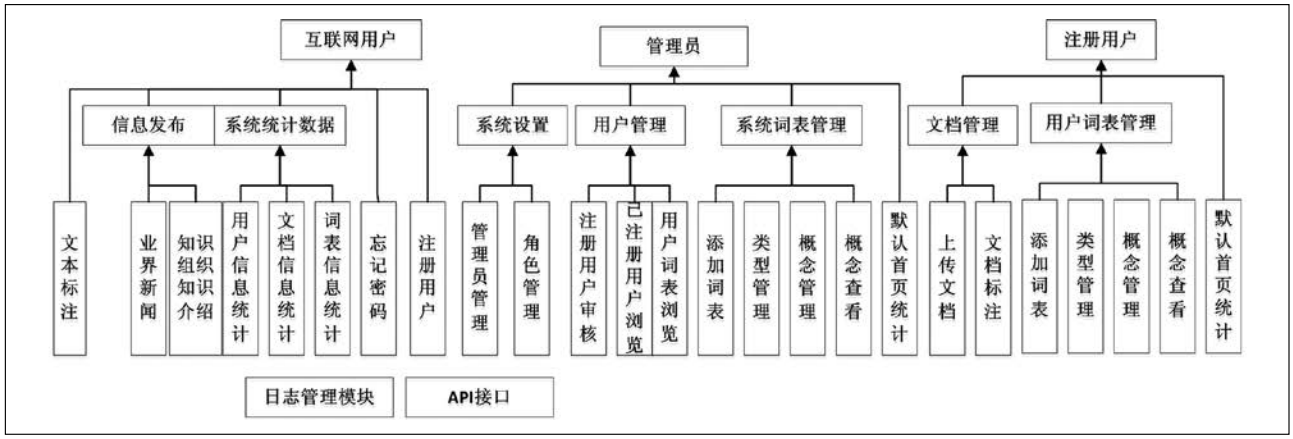


图1 系统功能结构图

切分标引法，即根据各种算法，在文献中去标注在叙词表中出现的概念。人工标注是一项繁重的脑力劳动[7]，它需要对标注的文献内容进行分析，再依据叙词表选择词汇对此文献的内容进行标注。但是，考虑到自动标注的有限性，目前尚不能完全代替人工标注。因此，本系统也增加了机器辅助标注，允许系统用户使用人工去标注。

## 2.1 多表联合标注系统整体设计

多表联合标注系统，有管理员、注册用户以及互联网用户等三类用户。管理员负责用户管理、系统词表管理，以及系统设置。注册用户主要有文档管理以及用户词表管理等功能。互联网用户相关功能有文本标注、注册登录，以及业界新闻浏览等。系统公共模块还包括日志管理模块及API接口等。多表联合标注系统详细功能结构如图1所示。

## 2.2 多表联合标注的流程

多表联合标注流程，主要有文档打开、词表设置、文本标注以及结果处理四个步骤。详细流程如图2所示。文档打开，是指用户可通过打开本地文档，或者粘贴来新建待标注文档。词表设置，是指添加或者移除用来标注的叙词表，并允许用户导入、编辑自己的词表，允许用户选择用来标注的词表、专业领域词表（如皮肤病、新能源汽车之类），以及通用的词表（如人员表、机构表）。文本标注，是指利用自动标注算法或者手工标注方式对原文进行标注。结果处理，包括标注效果预览、保存至本地以及标注结果复制和清空等功能。

以下将重点介绍文本标注。

文本标注包括自动标注与手工标注。其中，自动标注是指利用自动标注算法，对待标注文档利用包括的多部词表进行自动标注。如果文档原来标注过，则须先将原来的标注信息删除后，才能进行再次自动标注。手工标注，是指用户选择叙词表以及相应的类型属性，然后使用鼠标选择相关的待标注文本内容进行标签的创建、修改和删除。文本标注详细流程如图3所述。

标注后系统采用通过不同颜色显示来自不同词表的标签。待标注的文本内容通过嵌套标签来实现，这样，用户既能够保证不改变原文显示位置，又能让用户看到标注后的效果。例如，使用皮肤病词表，在待标注文档中标注“溃疡”。按照该词表，溃疡是一种“病症描述（PBD）”，其类型简称为PBD。这样，系统会标注如下：<stkos:c:PBD>溃疡</stkos:c:PBD>。其中，

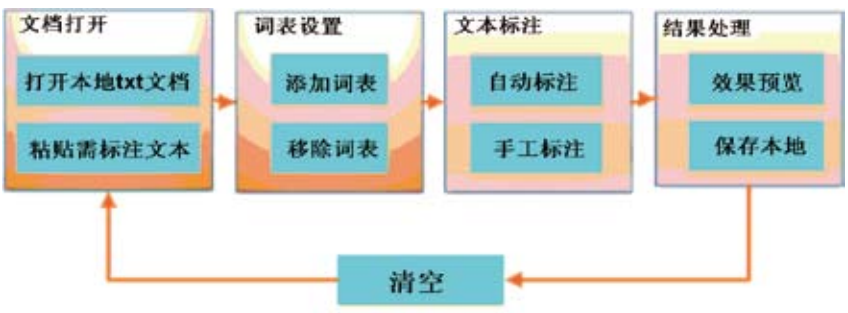


图2 多表联合标注流程

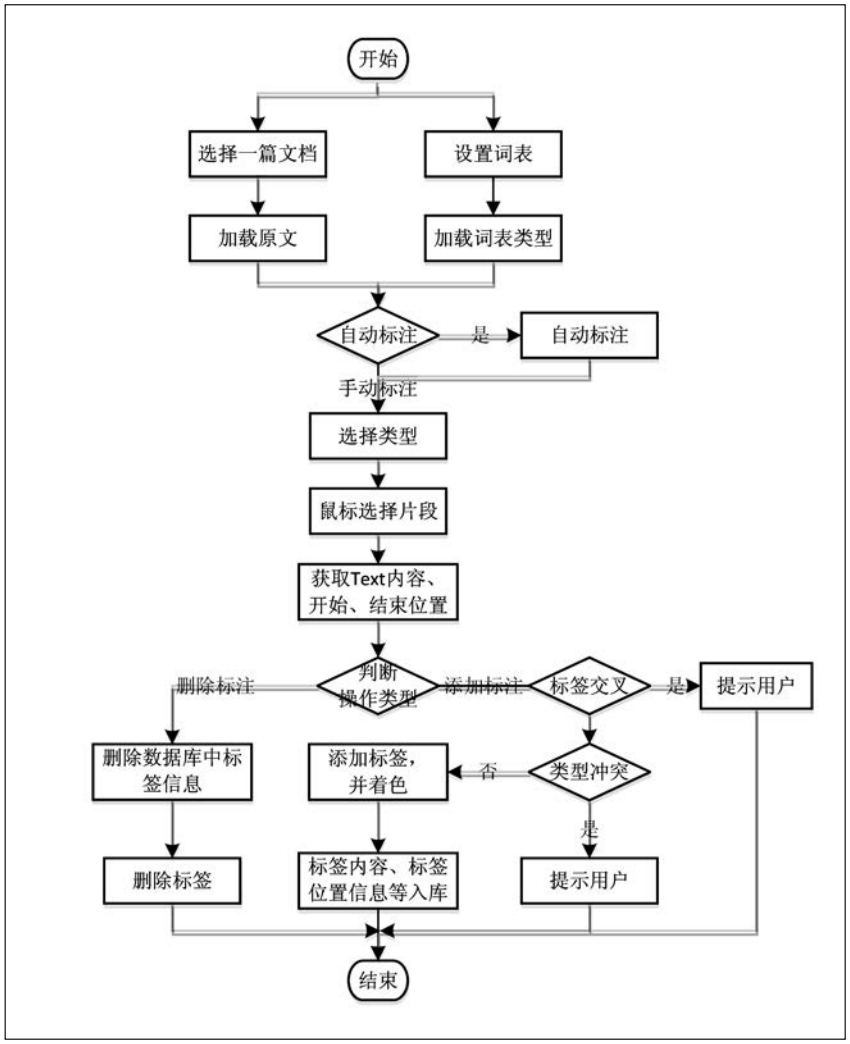


图3 标注详细流程图

stkos是指原文采用本系统创建的词表进行标注，c是指该用来标注的词表简称，PBD是指病症描述的代码。同时，系统还会在按照该词表规定的颜色等样式，显示标注后的效果，让用户所见即所得。

## 2.3 设计中应该注意的问题

多表联合标注系统涉及多部词表，同时，标注的标签较多，修改比较频繁。因此，在多表联合标注系统设计中，下列问题宜重点注意。

1) 文档管理

针对多用户平台，每个用户可能上传各自的文档，因此文档管理必不可少。系统在设计时，应该考虑到保持各个用户文档的相对独立性，同时支持通用的文档管理框架。针对此问题，系统应该编制相关的文档导入、导出规范。

2) 标注的形式

标注的形式主要两种，一是在原文中标注直接修改原文，二是将原文与标注分离。

直接修改原文，适合文本操作，标注位置容易确定，实现方案比较简单。但不足在于，由于文档额外空间不多，不好解决用户输入信息的填写位置。而且，一旦修改，所有位置都会出现偏移，会影响系统效率。

原文与标注分离，是指标注独立于原文存储，对同一原文可以显示不同人的标注。其优点是系统设计相对灵活，允许不同的用户对同一文档进行标注。同时，还方便对标注的管理，如增、删、改、查等。难点是记录标注位置，当原文变动时标注不容易精确指向。多数标注应用都把标注与原文分离，很多数Web应用使用原文与标注分离存储。

标注系统设计时，应该根据以上两种标注形式的情况决定，最终选择哪种标注方式，是选择在文中直接标注，还是原文与标注分离的形式。

3) 标注管理的灵活性

多表联合标注时，由于存在添加或者删除标注词表，因此标注的管理宜灵活。在灵活性方面，下文主要从增量标注和多视角展示标签两个方面进行分析。

增量标注。当用户增加一个新词表来标注已经标注过的文章，此时应该支持增量标注。实现增量标注，当用户新添加一个词表进行自动标注时，原来标注内容和标签不需变动，采用增量的方式进行自动标注（这样能够保证性能）。同样，删除已经标注的词表时，应该能够直接删除原标注词表的标签。

多视角展示标签。多表联合标注的目的是增加标注的维度，从而能够揭示原文的语义。因此，标注使用的标签，应该尽能够多视角展示。例如，能够通过词表分组的方式显示标签及内容，能够将某部词表的标注标签统计汇总等。

## 3 多表联合标注系统实现

多表联合标注系统采用B/S架构的软件平台，采用MyEclipse 9.0 + Tomcat 6为开发平台，采用SSH框架，JDK版本为1.6；数据库为MySQL 5.0。

以下以皮肤病领域下的文档标注为例，介绍其实现。多表联合标注系统标注页面如图4所示。页面上部是词表的管理区，中间为词表标签展示区，正中间是文档显示及标注区域，右侧是已标注标签的集中展示。

多表联合标注系统采用标注与原文分离的标注形式，在不改变原文的情况下，通过记录位置并将标签等信息存入数据库中。标注后的预览效果如图5所示。

在数据库中记录每个标注文本在原文中的开始位置和结束位置。系统导出按照相关规范进行导出。系统导出时可选择采用HTML还是XML格式，并定义了系统的标签以及固定格式。例如，为了与原文保存一致，系统导出时增加了原文的导出，其使用的标签名为“originalText”。HTML方式实际导出效果类似于图5所示，XML方式实际导出效果如图6所示（为节省篇章，原文部分只保留了部分内容）。


图4 多表联合标注界面

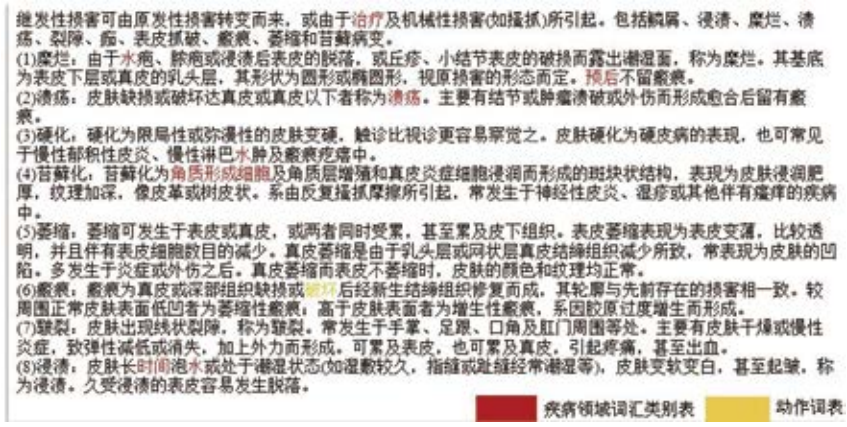


图5 标注后的预览效果


图6 标注后的文档导出

## 4 多表联合标注系统应用场景分析

### 4.1 深度检索系统

标注系统检索能够支持深度检索系统，其目的是实现专业化检索，以提供精确的检索结果。在现在信息泛滥的前提下，检索并不是缺少结果，而是缺少用户真正想要的结果。本标注系统检索提供精确的结果，应用于类似于QA这样的检索或者问答系统。

不同角色的人员针对相同的系统有不同的要求，不同的场景有不同的输入方式和习惯。针对医学信息方面的检索，角色可能包括患者和医务工作者两类。

（1）患者

患者在感觉身体不舒服后，登入系统，输入或选择不舒服的部位，输入或选择症状进行查询。页面示意如图7a所示。患者输入后系统自动联想，由用户选择后限定相关概念出现的词表。这样能够过滤一些并非该词表标注的内容，从而达到精确检索的目的。一般模式仅提供用户一个文本框，用户仅需输入关键字进行检索即可。用户在输入关键字时，如果关键字在多个类型中出现，那么提示用户检索哪个类型的数据。

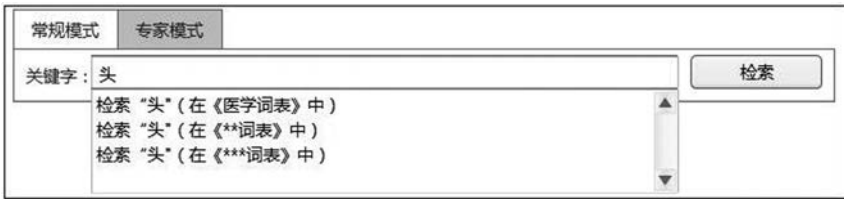


图7a 深度检索示意图（一）


图7b 深度检索示意图（二）

（2）医务专家

医务专家主要使用本系统进行资料检索。专家在检索时，检索信息比较具体、细致。医务专家除了输入部位+症状，还有可能输入词表中其他类型的数据。

专家模式提供用户选择词表、类型、概念进行检索，并可输入多组条件进行检索，在选择一个类型，不选择概念时，默认检索下属所有的类型，或者也可单独指定概念。这样能够支持“头”+“症状”、“部位”+“症状”方式的检索。检索示意如图7b所示。

当用户输入未确定概念，选择类型时，系统应根据已经标注的文档，检索类型相关的关键词，进行提示。例如用户输入了“部位”+“症状”，那么需要从已经标注的文档中检索部位和症状的所有匹配组合，显示出来，供用户选择。用户点击一个提示的组合，即可按点击的关键字进行检索，界面示意如图7b所示。

## 4.2 自定义标注引擎分析

用户自定义标注引擎，是指利用各个叙词表的特点，在原文上进行联合标注，形成组合的标注方案。用户自定义标注，并非标注词表全部类型属性，而是只标注用户感兴趣的某些特性，并且通过多部词表组合后形成综合方案。与自动标注不同的是，该标注引擎只标注用户感兴趣的内容，而自动标注使用的标签是某部词表的全部标签。例如：使用《药品》词表中的“副作用”标签可以标注所有药品相关的副作用。使用《皮肤病》词表中的“药品”名称可以标注治疗某皮肤病的所有药品。用户自定义“皮肤病药品可能带来的副作用”标注标签，则可以将两者结合起来，标注文档中出现的所有的皮肤病有关的药品名称及可能带来的副作用的内容。

自定义标注引擎还可以对标注的匹配规则、标注的算法等进行配置。匹配规则指定用户定制匹配的算法是最大匹配优先还是最小匹配优先。最大匹配是一次性尽可能标注多的内容，而最小匹配是尽可能标注细的内容。如“头痛”，针对最小匹配可能是<部位>头</部位><症状>痛</症状>。而最大匹配可能是<疾病>头痛</疾病>。

## 4.3 API功能

标注后导出的文档，还可以实现一些其他场景。例如：（1）提取标注的内容信息。利用API功能，可以把词表所有类型的全部内容或者部分内容依次取出来，从标注的文档中提取标注的内容信息。（2）提供词表服务信息。从标注的文档中，提供标签的解析服务，比如词表信息、标注的类型等信息。（3）标注后的统计信息。统计信息可能包括使用了几部词表、每部词表的类型数等。（4）操作信息。几部词表相联系的词条的联合操作，能够挖掘一些领域成果。

## 5 结语

本文介绍了叙词表多表联合标注系统设计和实现，主要讲述了多表联合标注系统的整体设计、标

注流程，以及设计中应该注意的问题。最后，探讨了多表联合标注系统可能存在的应用场景。文中对多表联合标注进行了一定程度的介绍。但是，多表联合标注是一个复杂的问题，例如各种标注算法的对比、标注指标的建立、标注效果的对比等都是实际系统设计与实现中要考虑的问题。

**参考文献**

[1] 周雪虹.制定文献编目著录细则若干问题的探讨[J].高校图书馆工作,2003,23(6) :42-43.
[2] 李素建,王厚峰.关键词自动标引的最大熵模型应用研究[J].计算机学报,2004,27(9):1192-1197.
[3] 章成志.基于条件随机场的自动标引模型研究[J].中国图书馆学报,2008(5):89-94.
[4] 程传鹏.微博自动标引关键技术的研究[J].计算机工程与应用,2011,47(34):137-140.
[5] 杨贺,杨奕虹,乔晓东,等.用于计算机辅助文献标引加工系统的自然语言词表构建[J].现代图书情报技术,2010(6):17-24.
[6] 朱嘉贤,白伟华,李吉桂.Web资源的多粒度语义标注及其应用技术研究[J].计算机科学,2011,38(8):83-87.
[7] 黄庆红.自动标引与机辅标引[J].现代图书情报技术,2002(S1):63,75.

**作者简介**

李鹏（1979-），硕士，助理研究员。研究方向：智能信息处理。E-mail: lipeng_cn@istic.ac.cn
朱礼军（1973-），博士，研究员。研究方向：智能信息处理。E-mail: zhulj@istic.ac.cn
辛之海（1975-），本科，工程师。研究方向：项目管理。E-mail: xinzhihai@outlook.com
刘亚洁（1988-），本科，工程师。研究方向：信息系统和软件工程。E-mail: liuyajie@istic.ac.cn

## Design and Implementation of Multi-Thesaurus Joint Tagging System

Li Peng, Zhu Lijun, Liu Yajie / Institute of Scientific and Technical Information of China, Beijing, 100038
Xin Zhihai / Open Sources Qizhi (Beijing) Software Co., Ltd., Beijing, 100125

Abstract: Universal thesaurus provides universal concept set while it has universality, coordination and compatibility features. And professional thesaurus with domain concepts focuses on professional areas. Document semantics can be revealed by tagging with concepts of multiple thesauri. This paper provided the design scheme of tagging system model with multiple thesauri, and introduced the function of text tagging module, as well as attention issues in system design. By discussing the documents' tagging in the skin disease field, the paper also summed up possible application scenarios about Multi-Thesaurus joint tagging. The system provides reference for mining document's meaning in the different perspective, and has laid the foundation for semantic retrieval.

Keywords: Thesaurus, Tagging, Semantics

（收稿日期：2013-10-09）